# Mining novel terpene synthases from large-scale sequence repositories

Čalounová Tereza1,2, Tajovská Adéla1, Smrčková Helena1,2, Bushuiev Roman1,3, Samusevich Raman1,4,5, Pluskal Tomáš1

*1 Institute of Organic Chemistry and Biochemistry CAS*

*2 Faculty of Science, Charles University in Prague*

*3 Faculty of Information Technology, Czech Technical University in Prague*

*4 Czech Institute of Informatics, Robotics and Cybernetics (CIIRC CTU)*

*3 Faculty of Chemical Technology, University of Chemistry and Technology, Prague*

Terpenoids are the largest and most diverse group of plant specialized metabolites with numerous medical or industrial applications. The terpenoid scaffolds are produced by enzymes called terpene synthases, which catalyze some of the most complex chemical reactions in biology.

We have created a database of characterized terpene synthases, which to this date contains more than one thousand entries, each including information about their taxonomy, terpene types and reactions. These terpene synthases were subjected to analysis of their domain architectures and structures, especially with respect to their taxonomy and terpene types. To perform the analysis of structures, we used AlphaFold 2 to obtain structure predictions for all terpene synthases in the database since there are very few experimentally determined structures of terpene synthases.

We mined UniParc, 1KP, Phytozome and TSA protein databases (606,214,611 protein sequences) using hidden Markov models from the Pfam database, which are associated with terpene synthases. From these databases we obtained 191,476 protein sequences of putative terpene synthases. We further generated a phylogenetic tree from both characterized and mined terpene synthase sequences. We have analyzed the phylogenetic tree and selected promising uncharacterized candidates based on the largest phylogenetic distance to the characterized sequences in the tree and their sequence reliability scores. The reliability score was calculated using the knowledge gained from the analysis of the database of characterized terpene synthases, primarily focusing on their sequence lengths and domain architectures.

Several selected candidate sequences will be experimentally characterized. Since we prioritize sequences distant from already characterized sequences, we presume that these sequences could synthesize novel terpene scaffolds.